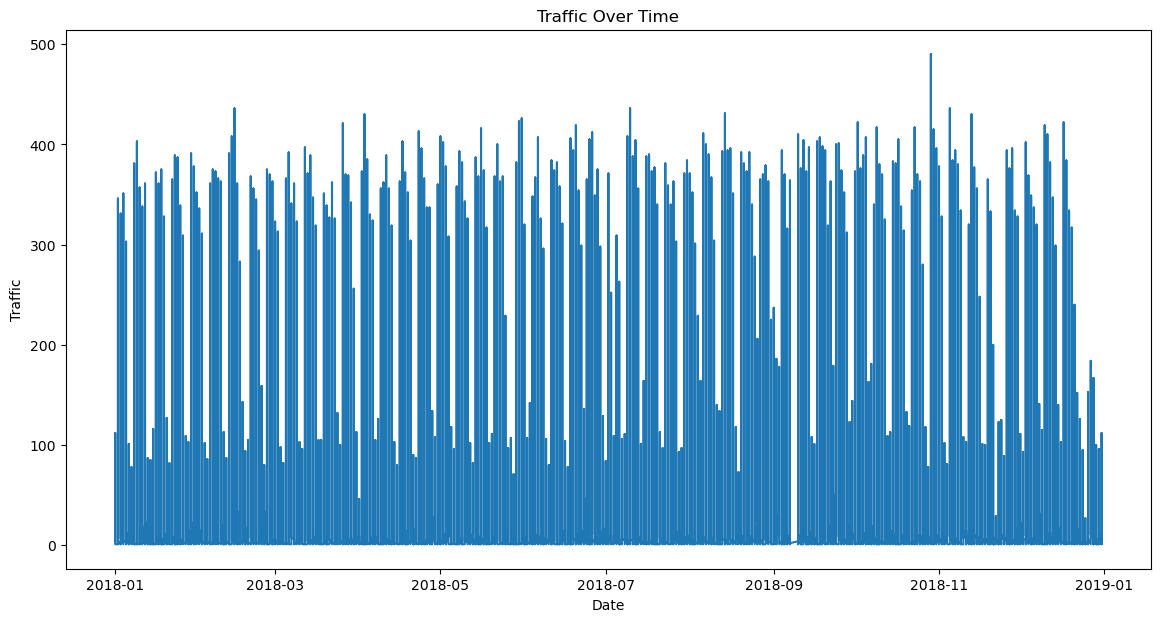
1. Introduction

This report presents a comprehensive analysis of the ridership patterns to West Oakland Station (WOAK) with an aim to understand the underlying trends, seasonal fluctuations, and unexplained variance in the number of passengers over time utilizing data from the Bay Area Rapid Transit system. This report employs time series and decomposition methods in order to extract meaningful insights that could aid in forecasting and operational planning.

1. Necessary Findings/Summary Reports/Conclusions/Graphs

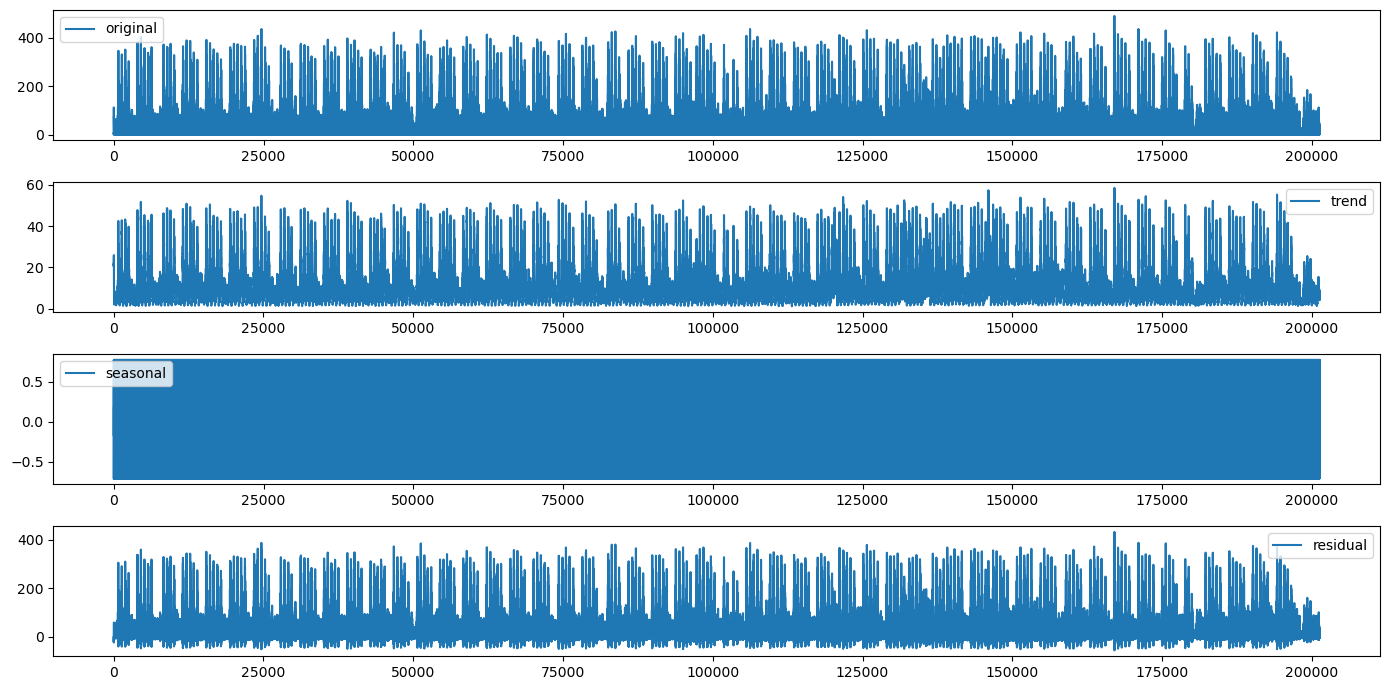
**Data Description**:  
Variables  
*Date  - date*  
*Hour  - hour of the day*  
*soo    - origin station*  
*dest   -  destination station*  
*n        - number of passengers*

| **Unnamed: 0** | **date** | **hour** | **soo** | **dest** | **n** |
| --- | --- | --- | --- | --- | --- |
| **0** | 1 | 2018-01-01 | 0 | 12TH | WOAK | 4 |
| **1** | 2 | 2018-01-01 | 0 | 16TH | WOAK | 5 |
| **2** | 3 | 2018-01-01 | 0 | 19TH | WOAK | 5 |
| **3** | 4 | 2018-01-01 | 0 | 24TH | WOAK | 5 |
| **4** | 5 | 2018-01-01 | 0 | CIVC | WOAK | 8 |



We plot the graph to find out how the traffic looks for the data range (from January 2018 to January 2019).

There is almost no traffic on Sept 10th? Why? There is very low traffic because maybe there is maintenance for the station causing no traffic to visit the West Oakland Station.

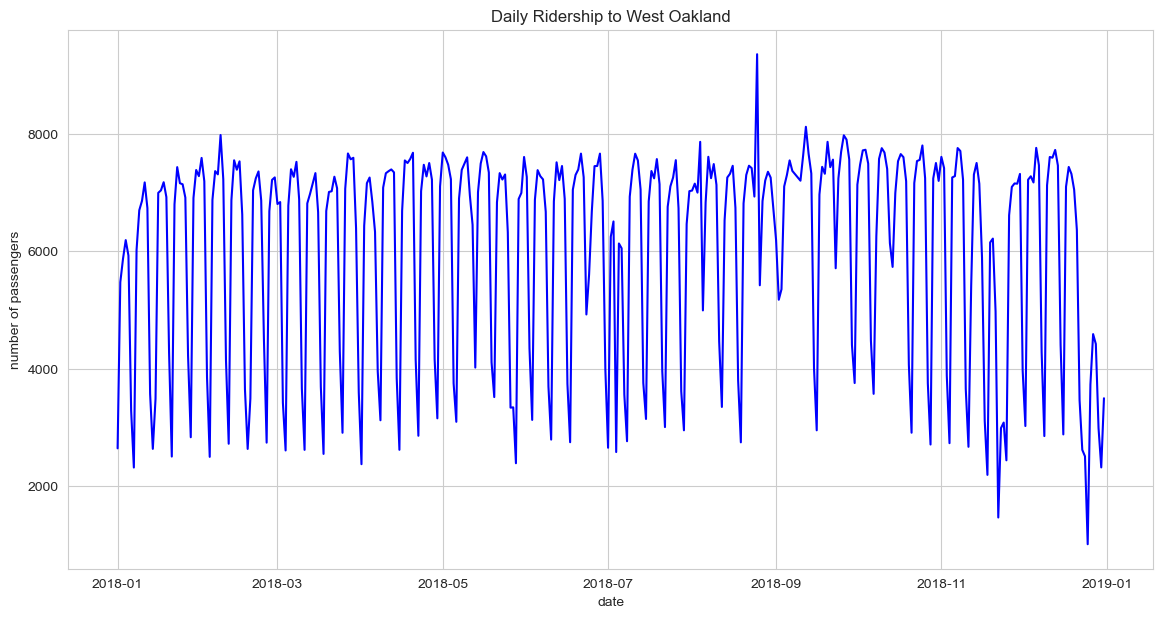


The first plot is the actual time series data and altered raw data that we are working with.

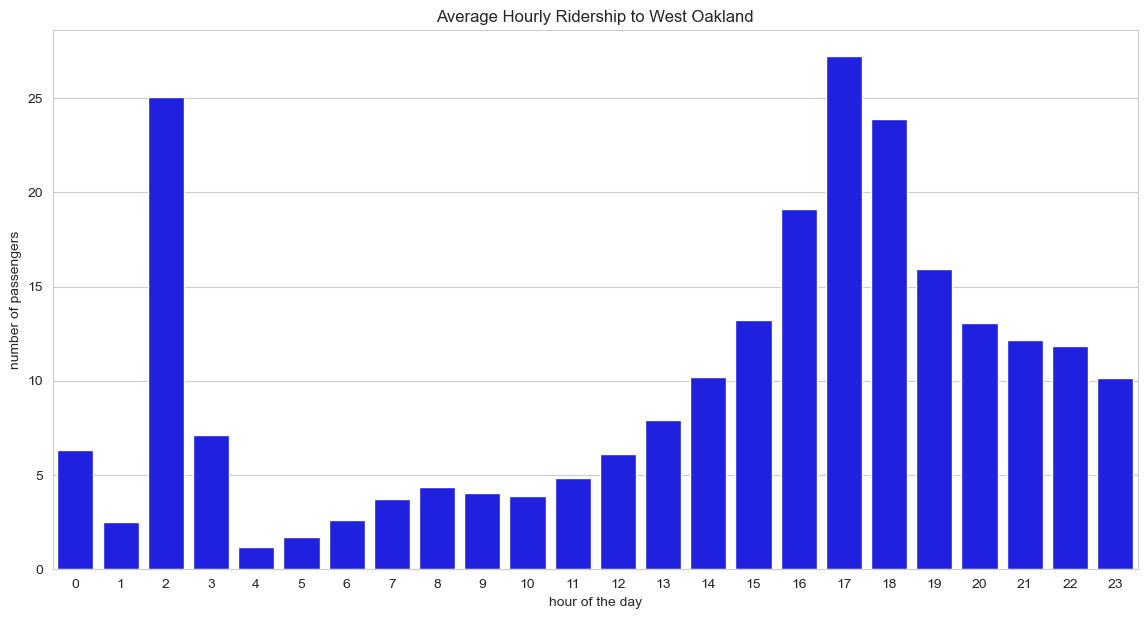
The second plot shows the trend which is a long-term movement in the time series. It indicates the general direction of the data over time.

The third plot has a seasonal component which reviews seasonal patterns in the data. Seasonality refers to fluctuations that occur at specific regular intervals of less than a year, such as monthly or quarterly.

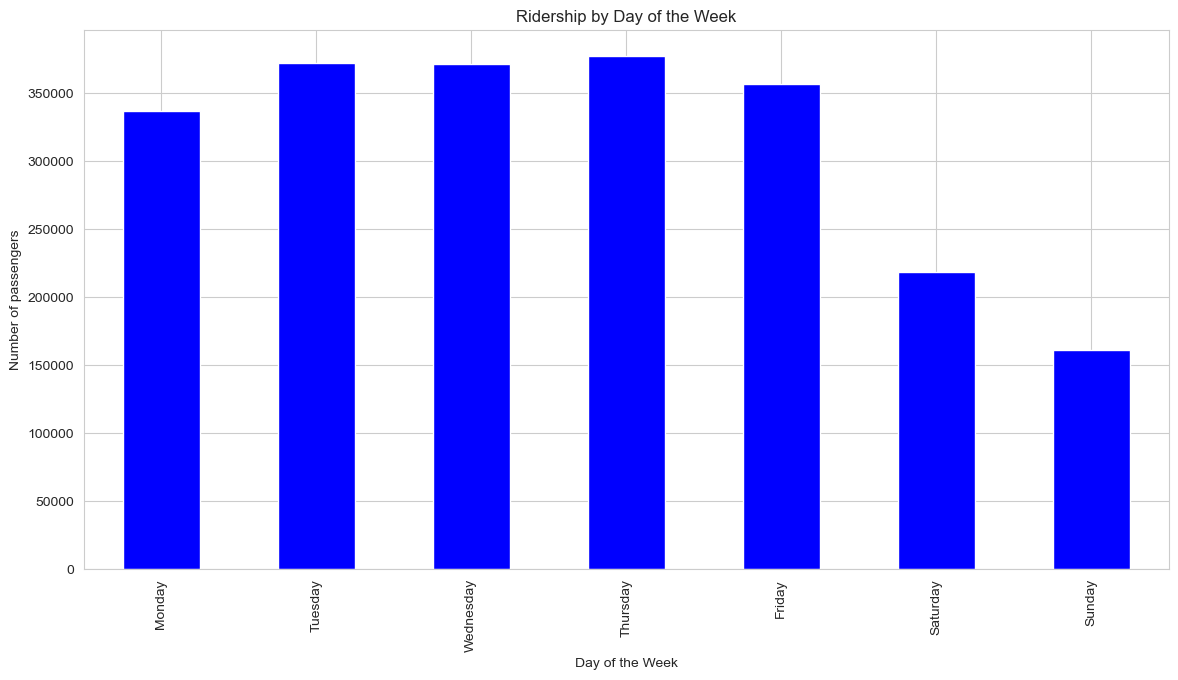
The fourth and last plot shows the residuals of the time series after the trend and seasonal component has been removed. What is left is random variation and cannot be explained by a seasonal or trend component.



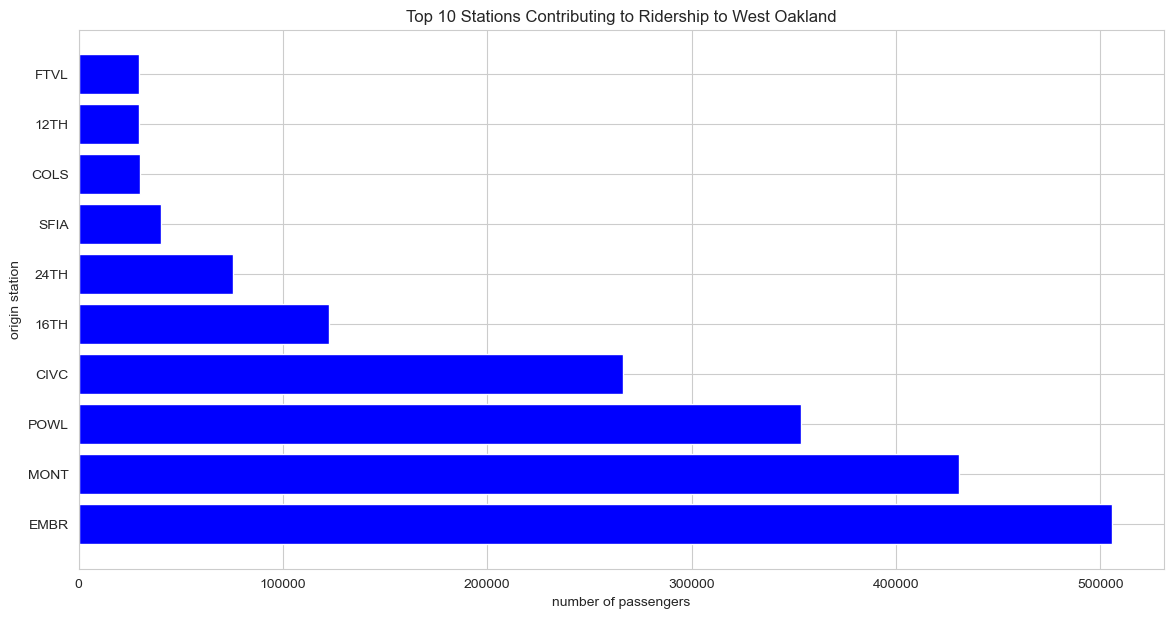
This displays the daily number of passengers that go to the West Oakland Station during the yearlong interval. There is a peak around September 2018 (August 15th?). It may indicate some major event happened for that period.



This plot indicates the average number of passengers for each hour. From this plot we can find out there is a peak time during 2am and 5pm. That may indicate many people work the night shift as well beside the normal afternoon rush.

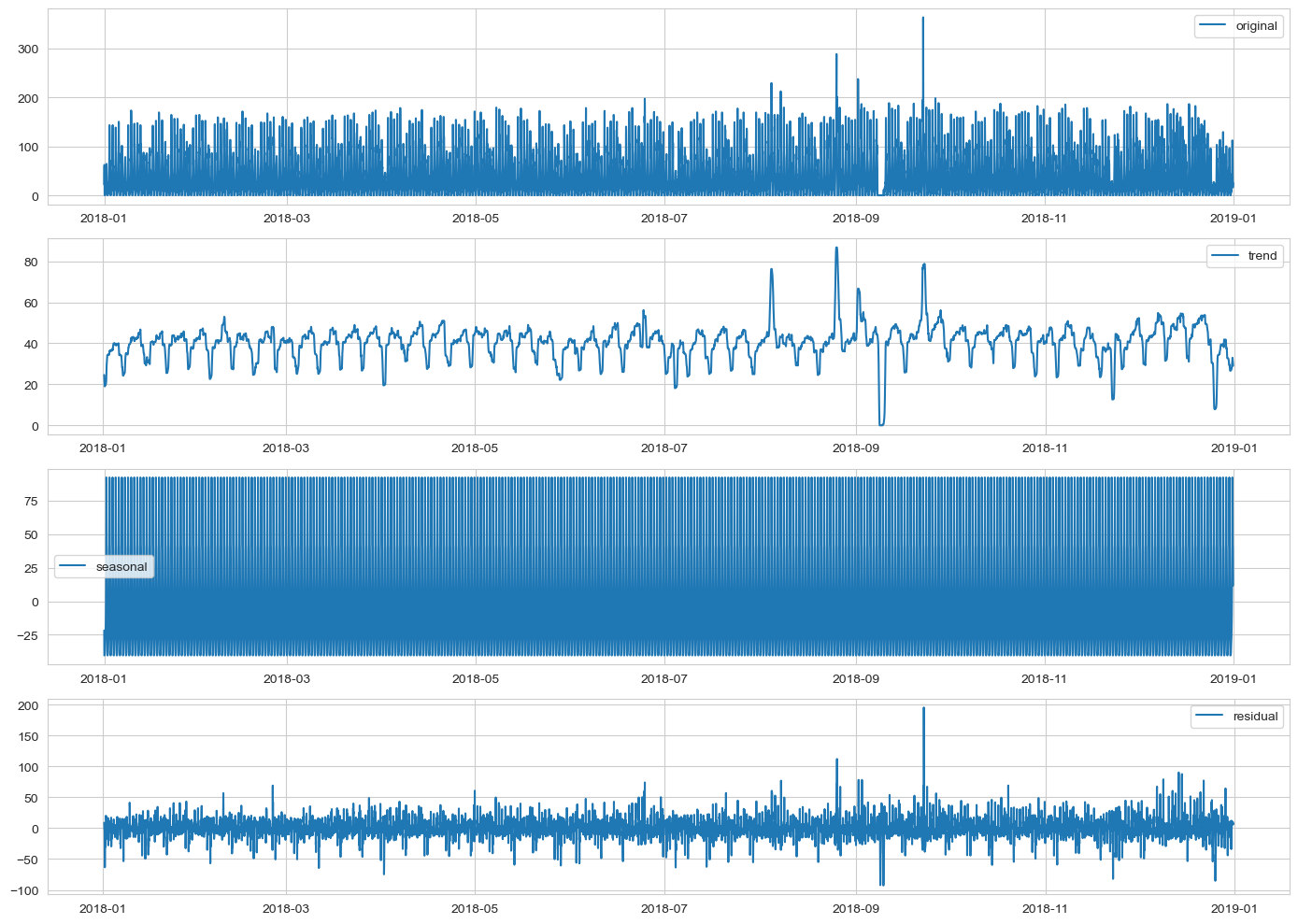


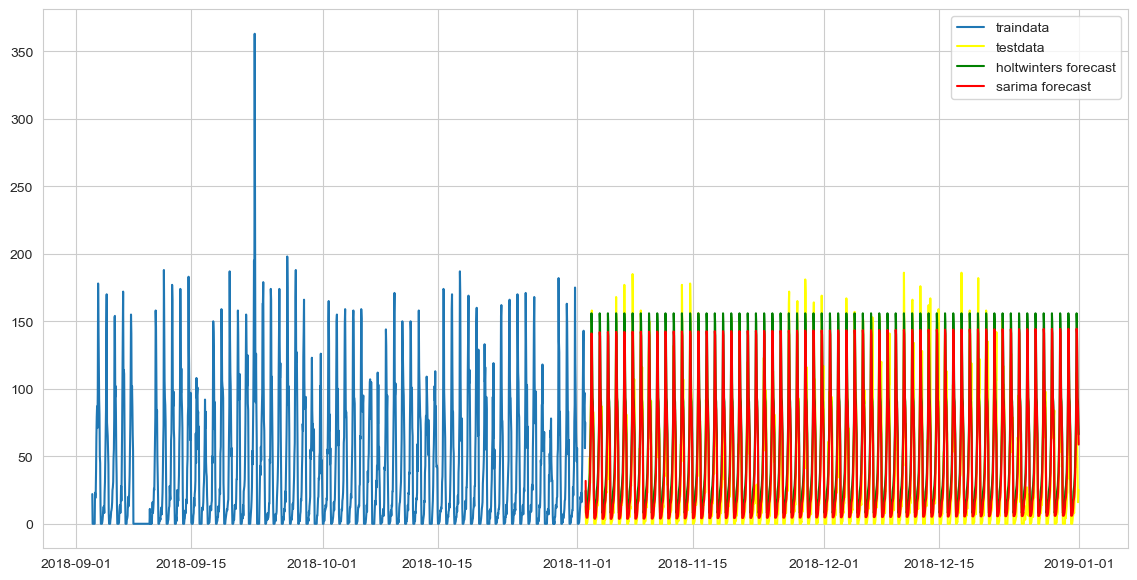
This plot indicates the number of passengers for each week day. During the weekend the traffic is slower compared to weekdays.



This plot indicates the top 10 stations which have the highest traffic travel to West Oakland station.

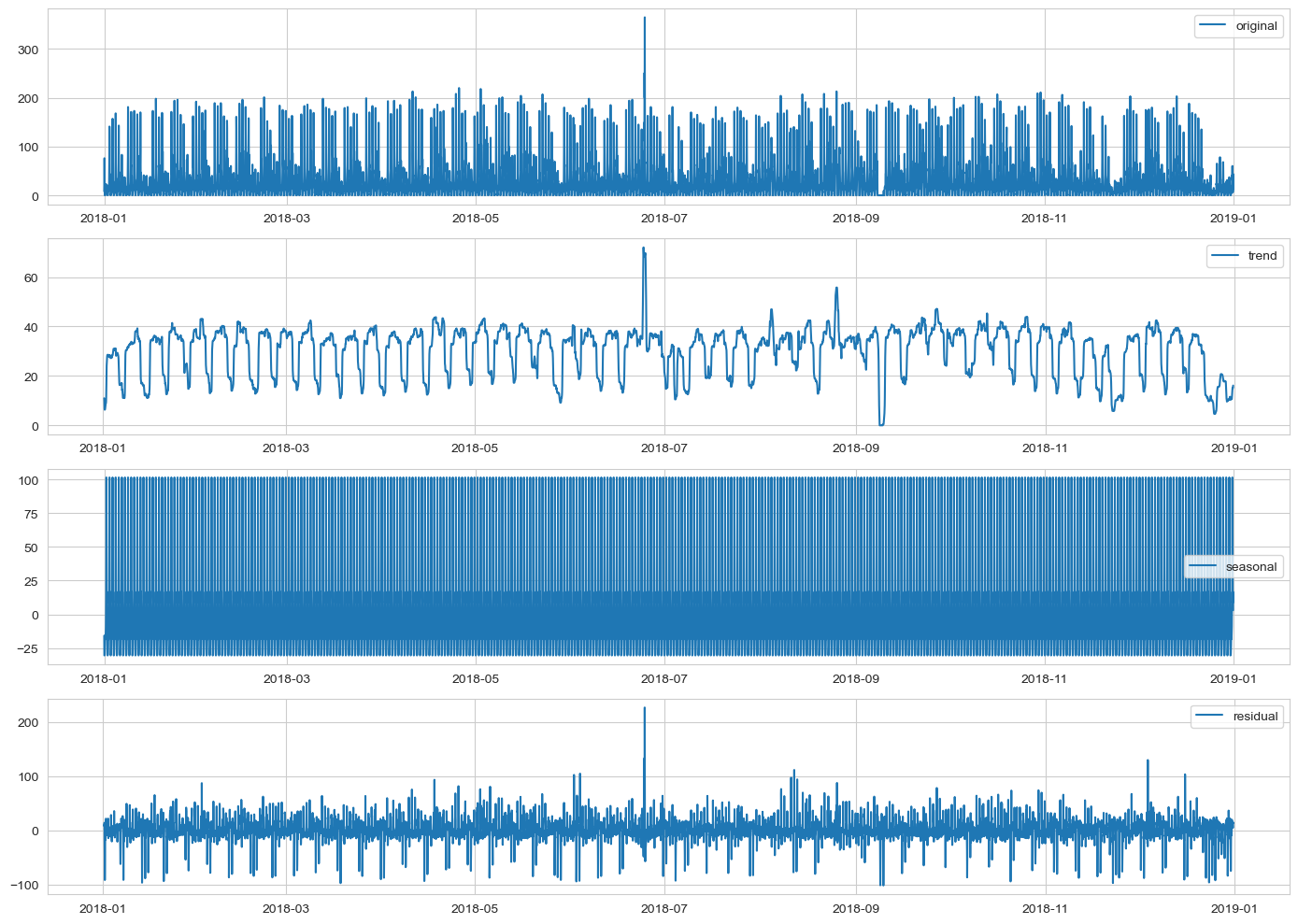
SARIMA

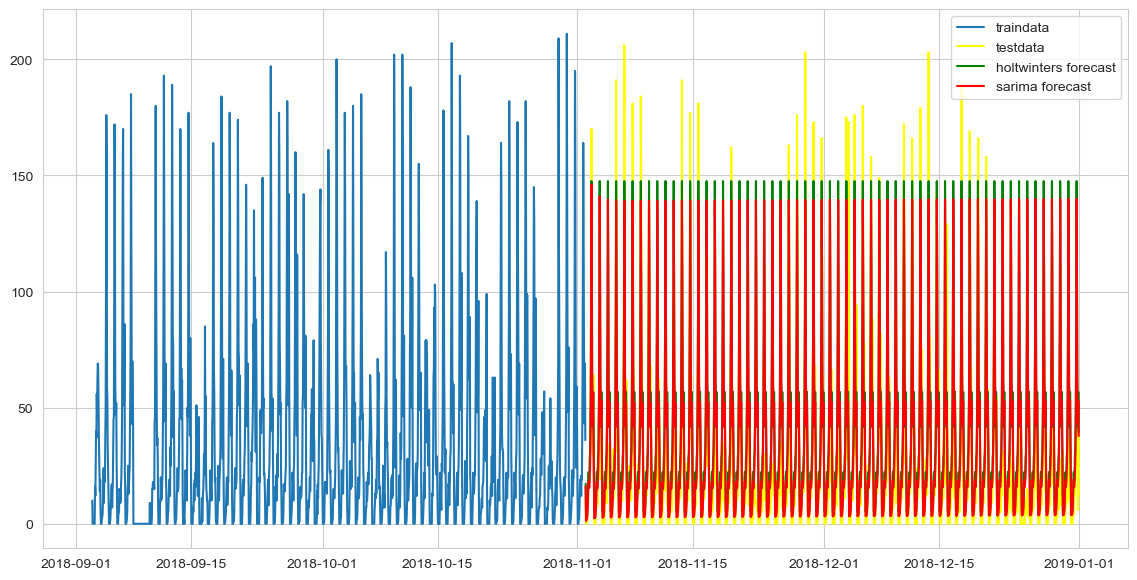




holtwinter mse: 575.6680069873984

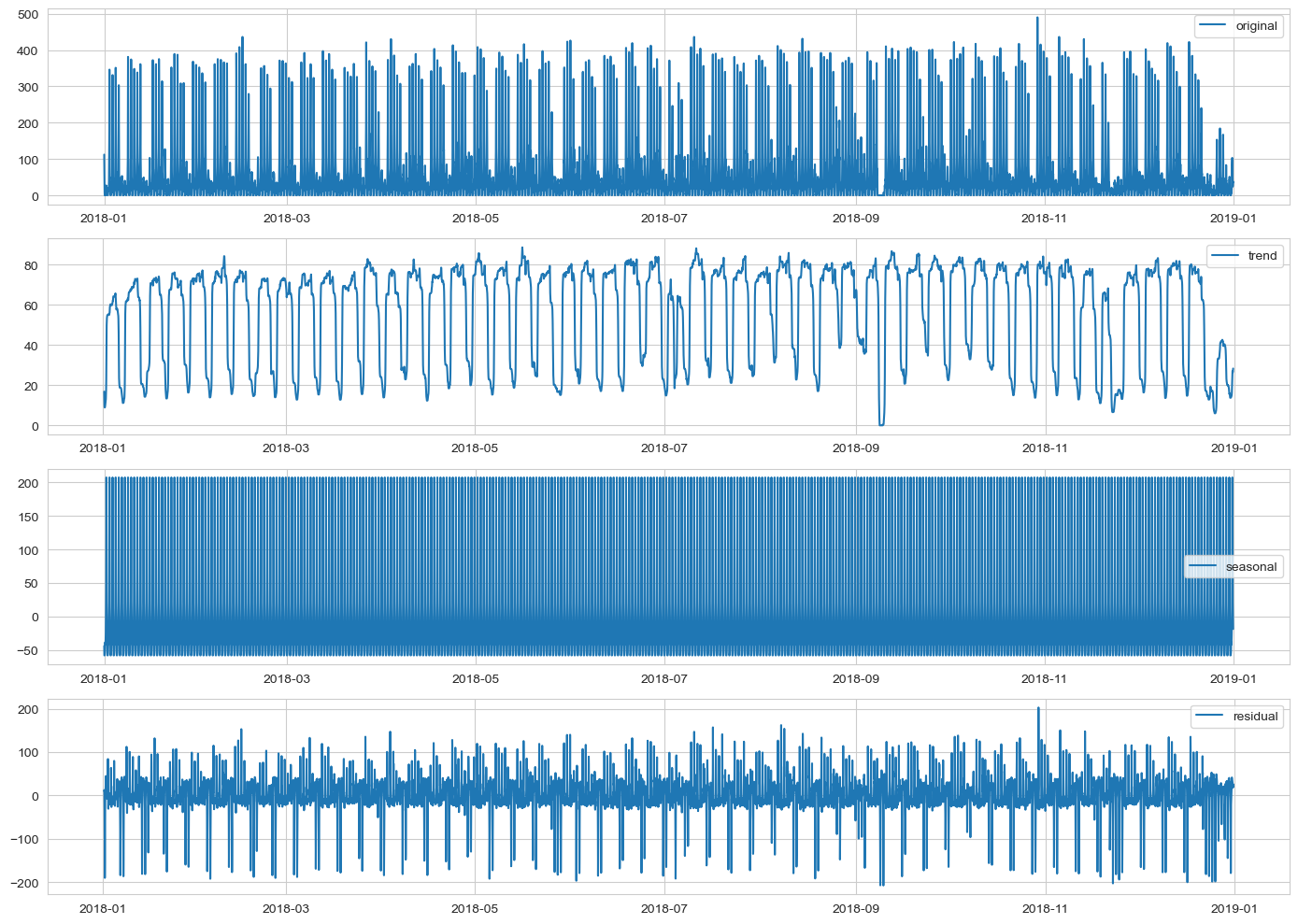
sarima mse: 412.0781105838061

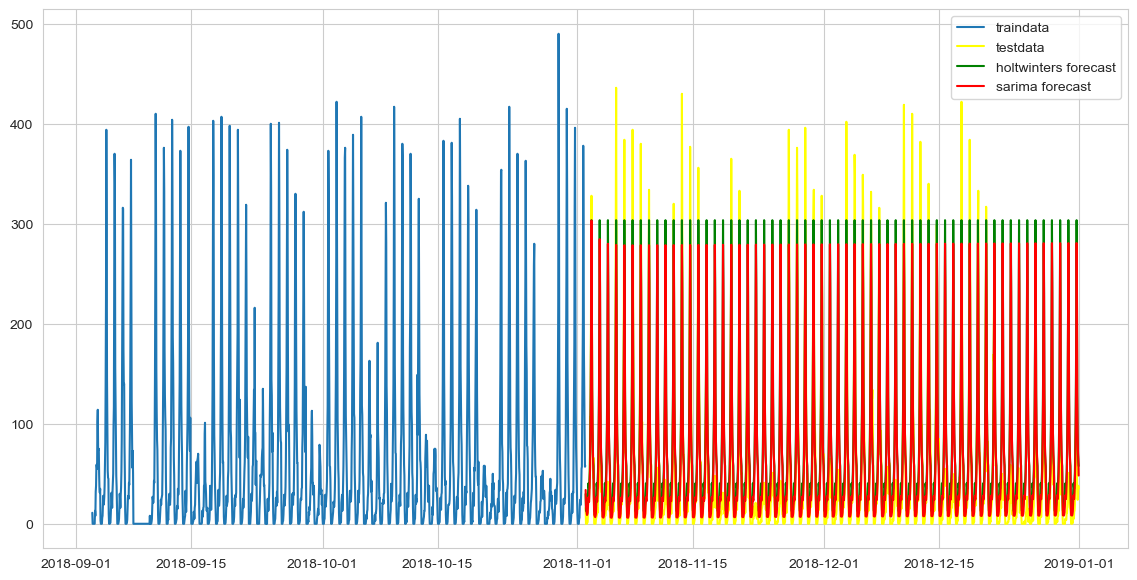




holtwinter mse: 699.7455736340945

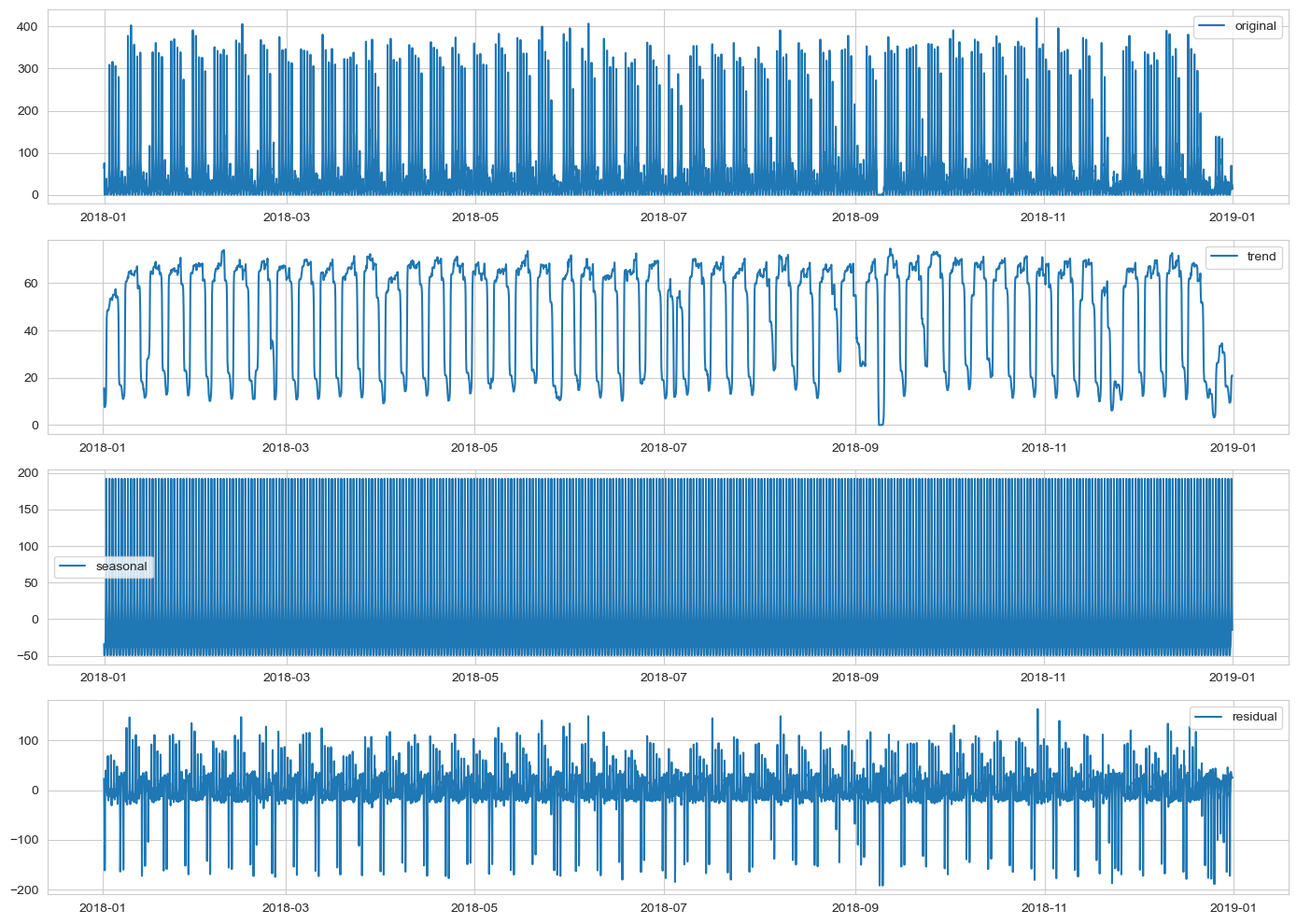
sarima mse: 615.7883488958937

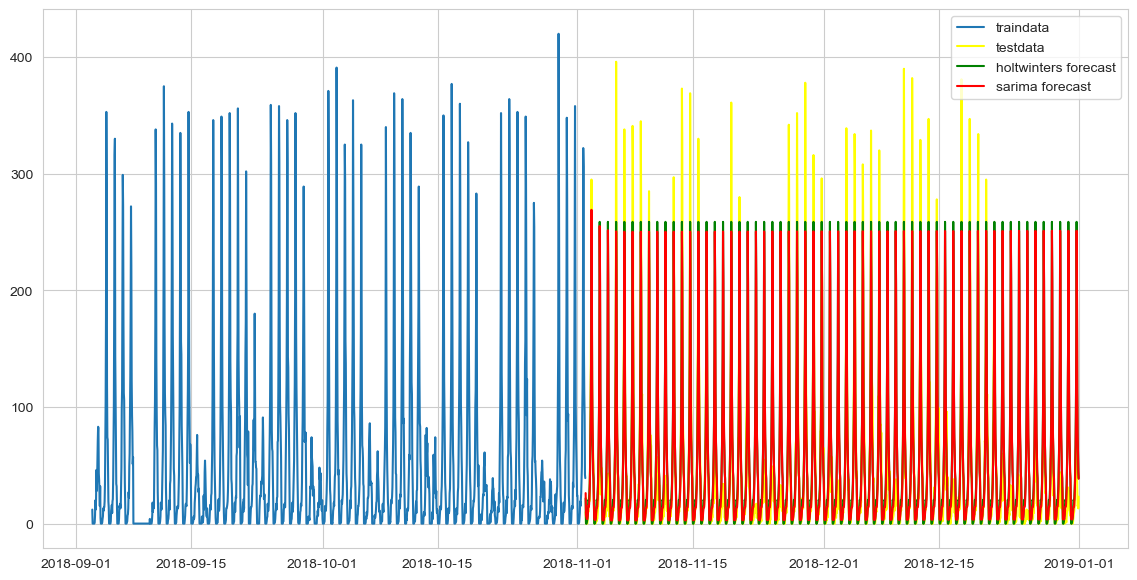




holtwinter mse: 3511.2888436212334

sarima mse: 2999.9147266004893





holtwinter mse: 2269.501804574054

sarima mse: 2222.51117083294

Explanation:

I did analysis on four stations: POWL, CIVC, EMBR, MONT:

The first plot is the actual time series data and altered raw data that we are working with.

The second plot shows the trend which is a long-term movement in the time series. It indicates the general direction of the data over time.

The third plot has a seasonal component which reviews seasonal patterns in the data. Seasonality refers to fluctuations that occur at specific regular intervals less than a year, such as monthly or quarterly.

The fourth and last plot shows the residuals of the time series after the trend and seasonal component has been removed. What is left is random variation and cannot be explained by a seasonal or trend component.

The fifth plot shows the split of data for test for training and testing purpose. It also uses two different forecast methods to predict the future traffic. From the graph it is hard to tell which forecast model is better, therefore, we calculate MSE to verify.

I calculated and printed out the Mean Squared Error(MSE) values for both the Holt Winter's model and the SARIMAX model. Holt Winters' forecast was defeated by the SARIMAX forecast in all 4 stations. A larger Holt Winters' MSE means that the model is more inaccurate compared to the SARIMAX MSE. Lower error (MSE) is good.

1. Submit supporting files of your work (a Python notebook for Python software users)